

Artificial Intelligence-based Automated International Classification of Diseases Coding: A Systematic Review

Abstract

Automated clinical coding, facilitated by artificial intelligence (AI) techniques like natural language processing and machine learning, has emerged as a promising approach to enhance coding efficiency and accuracy in healthcare. This review synthesizes current knowledge about AI-based automated coding of the International Classification of Diseases (ICD), with a focus on its challenges, benefits, and future research directions. Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, a systematic search was conducted across PubMed, Embase, Scopus, and Web of Science databases on January 1, 2024. Studies discussing challenges, advantages, and research gaps in AI-driven ICD coding were included. Out of 12,641 identified records, eight studies met the inclusion criteria. These studies highlighted six key challenges: extensive label space, imbalanced label distribution, lengthy documents, coding interpretability issues, ethical concerns, and lack of transparency. Ten major benefits of AI-based ICD coding were identified, including improved decision-making, data standardization, and increased coding accuracy. In addition, eight future directions were proposed, emphasizing interdisciplinary collaboration, transfer learning, transparency enhancement, and active learning techniques. Despite significant challenges, AI-based ICD coding holds substantial potential to revolutionize clinical coding by improving efficiency and accuracy. This review provides a comprehensive synthesis of current evidence and actionable insights for advancing research and practical implementation of automated ICD coding systems.

Keywords: *Artificial intelligence, autocoding, automatic coding, International Classification of Diseases*

Submitted: 29-Nov-2024

Revised: 10-Jan-2025

Accepted: 31-Jan-2025

Published: 06-Aug-2025

Introduction

The International Classification of Diseases (ICD), authorized by the World Health Organization, provides a standardized approach for categorizing medical diagnoses.^[1] This system, widely used in both clinical and research settings, has become integral to healthcare operations.^[2-5] ICD codes are instrumental in expediting administrative processes in hospitals and insurance companies, facilitating global data sharing, and supporting advanced statistical analyses.^[6]

ICD coding serves as a robust framework, with most hospital departments adhering to specific guidelines for documenting the diseases.^[7] Combining alphanumeric codes ensures consistency and enables effective communication across regions

and nations.^[8] Continuous improvements in disease classification systems have expanded the scope of covered illnesses, enhancing the relevance of ICD in medical practice.^[9,10]

The complexity of manual ICD coding, particularly with systems such as ICD-10-CM, which includes nearly 68,000 diagnostic codes, highlights the need for automation. The concept of automated clinical coding posits that computers can do clinical coding through artificial intelligence (AI) methods such as natural language processing (NLP) and machine learning (ML).^[11,12] Researchers have increasingly explored the potential of AI, including NLP and ML, to automate the coding process. These technologies promise to enhance accuracy, reduce time consumption, and mitigate human error in medical record management.^[13]

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Mousavi Baigi SF, Sarbaz M, Darroudi A, Kemmak FD, Aval RN, Kimiafar K. Artificial intelligence-based automated International Classification of Diseases coding: A systematic review. *J Med Signals Sens* 2025;15:22.

Seyyede Fatemeh Mousavi Baigi^{1,2}, Masoumeh Sarbaz¹, Ali Darroudi^{1,2}, Fatemeh Dahmardeh Kemmak^{1,2}, Reyhane Norouzi Aval^{1,2}, Khalil Kimiafar¹

¹Department of Health Information Technology, School of Paramedical and Rehabilitation Sciences, Mashhad University of Medical Sciences, Mashhad, Iran, ²Student Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran

Address for correspondence:
Dr. Khalil Kimiafar,
Department of Health Information Technology,
School of Paramedical and Rehabilitation Sciences,
Mashhad University of Medical Sciences, Mashhad, Iran.
E-mail: kimiafarkh@mums.ac.ir

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.jmss_76_24

Quick Response Code:



Recent studies indicate that there has been a growing body of publications on automated clinical coding using various AI methods, such as deep learning (DL) and NLP, over the past few years.^[1,14,15] Review articles offer valuable insights into the current state of knowledge within a specific field and suggest directions for future research.^[16] By synthesizing existing research and identifying gaps, literature reviews help advance the subject.^[17] These reviews, conducted in a clear, accurate, and reproducible manner, can guide future research in a particular area. Different types of review articles, such as systematic reviews, bibliometric analyses, and thematic reviews, provide fresh perspectives for future studies by reflecting on the current body of work in the field.^[18-20] In the area of AI-based automatic ICD coding in healthcare, several literature reviews have been published.^[21-27] This review synthesizes the existing body of the literature on AI-based automated ICD coding, with a focus on its applications, challenges, and future research opportunities. By analyzing prior studies, it aims to provide a comprehensive understanding of the field and identify the pathways for advancing this critical area of healthcare technology.

Instrument and Methods

Study design

To present the findings from the studies included in this systematic review, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. A comprehensive search was conducted on January 1, 2024, across the PubMed, Embase, Scopus, and Web of Science databases. The search utilized specific Medical Subject Headings (MeSH) and Emtree terms and keywords. The search strategy utilized a combination of MeSH and Emtree terms, including: ((Autocoding OR “automated coding” OR “automatic coding” OR “computer assisted coding” OR “automatic concept indexing” OR “computer coding” OR “automated extraction” OR “automatic extraction” OR “automated text mining” OR “automatic text mining”) OR (“clinical coding” OR “ICD” OR “international classification of diseases” OR “medical coding”) AND (“artificial intelligence” OR “computational intelligence” OR “computer reasoning” OR “machine intelligence” OR “deep learning” OR “computer-assisted diagnosis” OR “computer assisted diagnosis” OR “natural language processing”))).

Eligibility criteria

The review included studies that: (1) Focused specifically on AI-driven automated ICD coding; (2) Discussed challenges, advantages, or future research directions related to AI-based ICD coding; (3) Were peer-reviewed and published in English; (4) Were full-text accessible; and (5) Utilized systematic review or literature review methodologies.

Studies were excluded if they: (1) employed methodologies other than reviews (e.g., primary research studies); (2) focused on topics unrelated to AI-based ICD coding; (3)

were published in languages other than English; and (4) were conference abstracts or did not provide full-text access.

Methodological quality assessment

In this study, the Critical Appraisal Skills Programme (CASP) tool was utilized to assess the quality of the included studies.^[28] The CASP framework encompasses eight core criteria: clarity of study objectives, comprehensiveness of the literature review, methodological transparency, bias assessment, quality of data analysis, result transparency, practical applicability, and reporting of limitations. Each criterion was scored on a scale of 0 (poor), 1 (moderate), or 2 (excellent), yielding a maximum possible score of 16. Studies scoring below 9 were excluded due to insufficient methodological rigor, lack of comprehensiveness, or inadequate analysis quality. This threshold ensured the inclusion of studies with moderate-to-high quality, thereby enhancing the validity and reliability of the systematic review findings. This approach provided a robust foundation for synthesizing high-quality evidence.

Data extraction and synthesis

Titles and abstracts were independently assessed in accordance with the inclusion criteria, ensuring that no articles outside these requirements were included in the review. Two researchers then independently retrieved the full texts and evaluated them against the eligibility criteria. Any disagreements were resolved through discussion, with a third author serving as the arbitrator in cases where consensus could not be reached. Data extraction was carried out using a standardized form, which included the following data items: source (year and first author), study goals, study method, study results, challenges, benefits, and future directions of AI-based automatic ICD coding.

Findings

Study selection

The 12,641 papers that were discovered after the databases were searched are displayed in Figure 1. Following the removal of 6321 duplicate records, the titles and abstracts of 6320 studies were screened. After evaluating their relevance to the study’s objectives, 6190 studies were excluded. Subsequently, 130 articles were chosen for full-text review. In the end, eight studies met the inclusion criteria and were incorporated into this systematic review.

The quality assessment revealed that the majority of included studies demonstrated moderate to high methodological rigor, with clear objectives and comprehensive analyses. However, some studies lacked sufficient bias assessment and methodological transparency. The detailed evaluation scores are summarized in Table 1.

In general, eight eligible studies were included in this review. According to Table 2, of the eight included studies,

Table 1: Quality assessment scores of included studies using the critical appraisal skills programme checklist

Source (first author, year)	Clear aim	Comprehensive review	Methodology	Bias assessment	Quality of analysis	Transparency of results	Practical applications	Limitations	Total score (16)
Yan, 2022 ^[21]	2	2	1	0	2	2	2	0	11
Dong, 2022 ^[27]	2	1	1	0	2	2	2	1	11
Venkatesh, 2023 ^[24]	2	2	1	0	2	2	2	0	11
Stanfill, 2010 ^[25]	2	2	2	2	2	2	2	1	15
Wallace, 2023 ^[26]	2	1	1	0	2	2	2	1	11
Moons, 2020 ^[14]	2	1	1	0	2	2	2	1	11
Ji, 2022 ^[22]	2	1	1	0	2	2	2	0	10
Kaur, 2021 ^[15]	2	1	2	0	2	2	2	1	12

seven were reviews,^[14,15,21,22,24-26] and one study, in addition to the review, summarized the experience of clinical coding specialists in Scotland and the UK.^[27]

Tables 2 and 3 provide a comparative overview of the included studies, summarizing their characteristics, AI models, datasets, and evaluation metrics. This section complements these tables by highlighting the key insights from each study. Yan *et al.* explored the evolution of automated ICD coding, emphasizing three phases: rule-based, ML, and DL.^[21] Dong *et al.* highlighted the complementary nature of symbolic AI and neural network approaches, proposing their integration for optimal results.^[17] Venkatesh *et al.* identified the need for larger datasets and improved collaboration between developers and coders.^[24] Stanfill *et al.* categorized automated coding systems into those using predefined frameworks and those with custom classification schemes, noting their limited generalizability.^[25] Wallace *et al.* assessed CNN-based frameworks but called for more research to establish their superiority.^[26] Moons *et al.* examined hierarchical neural networks and demonstrated their advantages for large coding spaces.^[14] Ji *et al.* introduced an integrated framework combining text encoding and auxiliary information.^[22] Finally, Kaur *et al.* reviewed datasets and methods, highlighting the effectiveness of DL over ML for ICD coding.^[15]

Challenges

Overall, although the impact of AI-based automated ICD coding had many challenges, it was promising. In this study, common challenges in automatic ICD coding were identified. According to Figure 2, these challenges consist of six key items. The following list explains each identified challenge in detail: (1) large label space;^[21] (2) unbalanced label distribution;^[15,21,22] (3) long text of documents;^[15,21,27] (4) poor interpretability of coding;^[15,21] (5) moral and social consequences;^[24] and (6) lack of explainability and transparency.^[24]

Advantages

Despite the many challenges, AI-based automatic ICD coding has potential benefits that have attracted the efforts of policymakers and developers in this field today. The following is a list of these benefits: (1) helping experts

in decision-making;^[21] (2) disease and death statistics and analysis;^[21] (3) standardization and sharing of medical data;^[21] (4) payment of medical expenses;^[21,25] (5) medical record management;^[21] (6) hospital evaluation management;^[21] (7) increasing the speed of automatic coding compared to manual coding;^[27] (8) coding based on a standard that is nonbias and prevents mental coding;^[27] (9) enhancing the precision, caliber, and effectiveness of manual coding;^[25-27] and (10) active learning in unlabeled data classification.^[15]

Future directions

Of the 8 included studies, 2 specifically suggested future directions for AI-based automatic ICD coding.^[15,24] After examining the challenges of AI in automatic ICD coding, eight critical factors have been identified as priorities for future research: (a) interdisciplinary cooperation,^[14,15,24] (b) reliable data sources,^[15] (c) transparency in coding processes,^[14] (d) exclusive coding for customized research,^[24] (e) national acceptance and coding regulations,^[15] (f) addressing complex problems,^[15,24,25] (g) adopting transfer learning approaches for automatic ICD coding,^[15] and (h) Methods of active learning and reinforcement learning for clinical categorization issues.^[15]

Discussion

In this review, 8 studies were identified that met all the listed inclusion criteria. Of the eight included studies, seven were reviews,^[14,15,21,22,24-26] and one study, in addition to the review, summarized the experience of clinical coding specialists in Scotland and the UK.^[27]

Overall, although the impact of AI-based automated ICD coding had many challenges, it was promising. In this study, common challenges in automatic ICD coding were identified. It consisted of six items: (1) large label space;^[21] (2) unbalanced label distribution;^[15,21,22] (3) long text of documents;^[15,21,27] (4) poor interpretability of coding;^[15,21] (5) moral and social consequences,^[24] and (6) lack of explainability and transparency.^[24]

Large label space

medical codes with many dimensions. The fact that medical notes are linked to several diseases is typically seen as a serious multi-label categorization issue with a sizable label set. For instance, ICD-9-CM includes approximately 13,500

diagnostic codes and 4,000 procedure codes, each containing up to four digits. In its successor, ICD-10-CM, the number of diagnostic codes will exceed 70,000, while procedure codes will total 72,000, with each code extending to seven digits. The vast coding space presents challenges in prediction, and these issues are expected to grow as more iterations of ICD are introduced. In addition, code sets like ICD-10 are evolving rapidly, and training datasets for these codes remain incomplete, while ACC models have yet to fully grasp the logic and rules underlying coding decisions.^[21]

Unbalanced label distribution

the scenario in actual medical settings is consistent with the extremely frequent occurrence of a limited number of codes. The majority of hospital patients have common illnesses.

Many uncommon diseases in the ICD classification may never manifest, while a few other disorders may be found on occasion. Furthermore, some ICD codes that were just added in the most current edition might not be used at all or only infrequently.^[21] However, individuals with complex disorders are linked to dozens of codes in the entire coding space, whereas a patient is often diagnosed with only a few codes.^[15,21]

Long text of documents

in general, medical records can be quite lengthy.^[15,21,27] This is particularly true for electronic medical records (EMRs), which employ a lot of complicated medical and professional words and contain information about past and present medical history, exam results, diagnoses, and more.

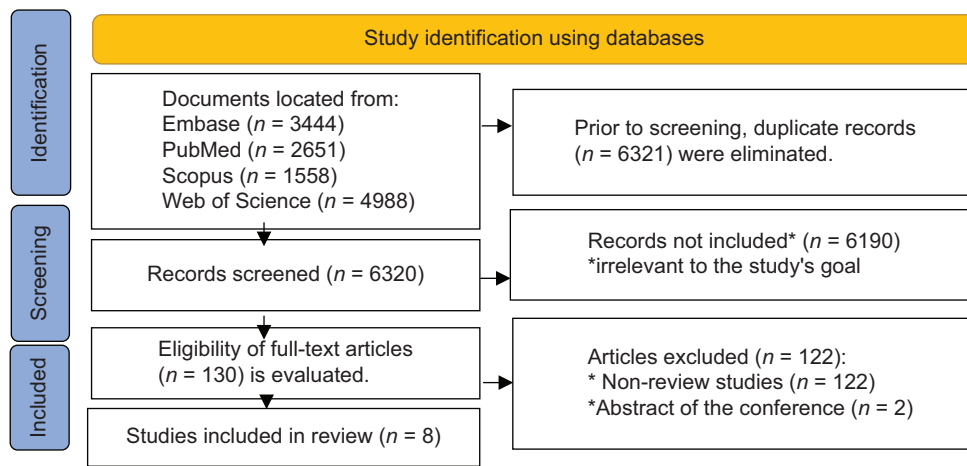


Figure 1: Flowchart of the study selection



Figure 2: Challenges, benefits and future directions in artificial intelligence-based automatic International Classification of Diseases coding. ICD: International Classification of Diseases. AI: Artificial intelligence

Table 2: Summary of key characteristics of included studies

Source (first author, year)	Study method	Study goals	Study results
Yan, 2022 ^[21]	A review	Examine the evolution of automatic ICD coding over the past few decades, starting with rule-based, conventional ML and ending with neural network-based techniques	The evolution of automatic ICD coding was divided into three phases 1. Rule-based phase: Focused on replacing manual labor by extracting coding rules and implementing them in logic systems 2. Conventional ML phase: Used techniques like SVM to predict ICD codes, emphasizing feature development and classification 3. Deep learning phase: Introduced advanced neural network models to improve the representation and prediction of health-related documents
Dong, 2022 ^[27]	A cross sectional study with review	To explore the concept of automated clinical coding, identify obstacles from AI and NLP perspectives, and evaluate shortcomings in current DL-based methodologies	AI shows great potential for automated clinical coding despite significant organizational and technical challenges. Overcoming these barriers in the next 5 years could lead to the successful development and implementation of AI-based systems for clinical coding
Venkatesh, 2023 ^[24]	A review	To analyze the challenges and potential solutions for implementing a successful automated healthcare coding system	Automated clinical coding technology requires significant innovation and adoption milestones. Future developments, driven by cooperation among developers, new training data, and advancements in AI and NLP, will revolutionize coding practices, revenue management, and invoicing through both assisted and autonomous models
Stanfill, 2010 ^[25]	A systematic literature review	To evaluate the effectiveness of various automated coding and classification methods through a comprehensive literature review	This study highlights the potential of automated coding systems but emphasizes the need for field-specific considerations before implementation. Key areas requiring further research include performance standards for real-world clinical activities, such as data delivery for decision support, clinical research, and quality measurement
Wallace, 2023 ^[26]	A rapid review	To evaluate the current applications of CNNs for predicting ICD codes from EMRs	The study highlights the potential of CNN frameworks but notes ongoing debates about their optimality for ICD coding. Preliminary findings emphasize enhancement techniques such as word embedding and neural transfer learning. However, further research is required to validate these methods and establish the superior performance of CNNs, as current evidence remains limited
Moons, 2020 ^[14]	A review	To review the use of deep neural networks for automatically classifying clinical data into ICD codes	This study evaluated neural network-based methods for ICD coding and highlighted the importance of loss functions for handling long descriptions. Models with hierarchical objectives were found particularly useful when training data was limited. However, increasing dataset size with noisy medical records negatively impacted model performance
Ji, 2022 ^[22]	A review	To outline current state-of-the-art models and provide an integrated framework for understanding the fundamentals of medical coding	This review presents a unified framework for medical coding, consisting of four key components: Deep encoder architectures, decoder modules for mapping hidden representations to medical codes, encoder modules for extracting text features, and the utilization of auxiliary data for improved performance

Contd...

Table 2: Contd...

Source (first author, year)	Study method	Study goals	Study results
Kaur, 2021 ^[15]	A systematic literature review	To provide a comprehensive overview of an automated clinical coding system that assigns ICD codes to discharge summaries using NLP, ML, and DL approaches	The review identified two public datasets and four hospital-acquired datasets for automatic coding. It highlights the significant improvements achieved by DL models compared to ML methods, employing 14 NLP techniques alongside feature extraction and embedding strategies. Various assessment criteria were utilized to evaluate the effectiveness of classification techniques

CNN – Convolutional neural networks; NLP – Natural language processing; ML – Machine learning; DL – Deep learning; ICD – International Classification of Diseases; SVM – Support vector machine; AI – Artificial intelligence; EMR – Electronic medical records

Furthermore, medical records may contain a large number of nicknames, acronyms, and uncommon keywords in addition to a few misspelled words. As a result, EMRs display text that is very dispersed and noisy. As a result, inaccuracies or redundant information will always impair the modeling of lengthy medical texts, possibly resulting in the loss of crucial information. As far as we are aware, not many approaches have been put forth to address this issue. To enhance the presentation of lengthy texts, some unique procedures may be employed. The self-distillation learning method was employed by Pascual *et al.*^[29] to address the issue of noise in lengthy texts. Their student model learned how to extract important information from a lengthy text by using the actual text with noise, while their teacher model used the description of the target code. Lu *et al.*^[30] used attention to compress the features that were produced via filters. To identify significant neighboring phrases and eliminate noise, the sliding windows of the filters must be combined because they produce extra information. Although these initiatives point in some potential avenues for addressing this issue, more work remains until long document modeling is achieved to a satisfying degree.

Limited interpretability in coding

A major limitation of DL models is their lack of interpretability. However, for automated ICD coding to be effective in clinical environments, it must be sufficiently interpretable to assist in decision-making. In other words, incorporating relevant contextual information and decision support during ICD code predictions is essential for enhancing the model's interpretability.

Moral and social consequences

For patients, providers, payers, and society at large, AI for medical coding may have moral and societal ramifications. For example, AI models may introduce biases or errors in medical coding that could affect the quality of care, reimbursement rates, or patient health outcomes. Furthermore, AI models may replace or reduce the role of human coders, which could affect their job satisfaction, skill development, or career prospects.^[24]

Lack of explainability and transparency

AI for medical coding needs to provide clear and comprehensible explanations for its coding decisions,

especially when they differ from human coders or when challenged by auditors or regulators. Nevertheless, AI models tend to be intricate and lacking transparency, which makes it challenging to comprehend their internal processes or the reasoning behind specific decisions. Therefore, AI models may lack the explainability and transparency necessary for trust and accountability in medical coding.^[24]

Despite the many challenges, AI-based automatic ICD coding has potential benefits that have attracted the efforts of policymakers and developers in this field today. Below is a list of these benefits: (1) assisting healthcare professionals in decision-making;^[21] (2) analyzing disease patterns and mortality rates;^[21] (3) facilitating the standardization and sharing of healthcare data;^[21] (4) managing medical billing and expenses;^[21,25] (5) organizing and maintaining medical records;^[21] (6) managing hospital performance evaluations;^[21] (7) increasing the speed of automatic coding compared to manual coding;^[27] (8) coding based on a standard that is nonbias and prevents mental coding;^[27] (9) improving the accuracy, quality, and efficiency of manual coding;^[25-27] and (10) active learning in unlabeled data classification.^[15] To further illustrate these potential benefits, a recent study by Dai *et al.*^[31] provides a real-world example of AI implementation in ICD coding within a hospital environment. Conducted at Kaohsiung Medical University Chung-Ho Memorial Hospital, the study evaluated an NLP-driven AI-assisted ICD-10-CM coding system integrated into the workflow of certified coding specialists. The system demonstrated strong performance, with the GPT-2 model achieving an F1-score of 0.667 on test data and 0.621 on the real hospital data. Moreover, the model showed high agreement with coding specialists (Cohen's $\kappa = 0.714$), significantly outperforming traditional models. This study highlights the potential of AI-assisted systems to enhance coding efficiency and reduce manual workload in the real-world settings, while emphasizing the importance of further research on scalability and data integration for broader adoption. This example not only demonstrates the practicality of AI in automated ICD coding but also sets the stage for exploring its broader applications in healthcare.

Beyond automated ICD coding, AI has demonstrated significant potential in other areas of healthcare. For instance, its application in occupational therapy has

Table 3: Comparative overview of artificial intelligence models, datasets, and evaluation metrics in included studies

Source (first author, year)	AI model	Dataset	Evaluation metrics
Yan, 2022 ^[21]	Rule-based, SVM, CNN, RNN, GNN, pretrained models (BioBERT, PubMedBERT)	MIMIC-II, MIMIC-III, MIMIC-IV, CCHMC, UKLarge, UKSmall, Xiangya, CN-Full	Precision, recall, F1-score (micro/macro), AUC, P@K
Dong, 2022 ^[27]	Rule-based, deep learning (CNN, RNN, BERT-based models), knowledge-based AI	MIMIC-III, real-world UK data (mixed healthcare sources)	Accuracy, precision, recall, micro-F1, macro-F1
Venkatesh, 2023 ^[24]	Rule-based, NLP, CNN	MIMIC-III, synthetic datasets from hospitals	Accuracy, F1-score, recall, transparency
Stanfill, 2010 ^[25]	Rule-based, NLP-based systems	Multiple datasets (SNOMED, ICD-9-CM, UMLS)	Recall, precision, sensitivity, specificity
Wallace, 2023 ^[26]	CNN frameworks (e.g., MultiResCNN, CAML, DR-CAML)	MIMIC-II, MIMIC-III, Chinese EMRs	Accuracy, macro-AUC, micro-AUC, F1-score
Moons, 2020 ^[14]	CNN, BiGRU, DR-CAML, MVC-RLDA	MIMIC-III, CodiEsp	Micro-F1, macro-F1, micro-AUC, P@5
Ji, 2022 ^[22]	Encoder-decoder framework (e.g., CNN, RNN, transformers, GNNs)	MIMIC-III, ICD datasets, real-world EHR datasets	Accuracy, precision, recall, F1-score, AUC
Kaur, 2021 ^[15]	NLP, ML, DL	6 datasets: 2 public (e.g., MIMIC-III), 4 hospital-acquired datasets	Accuracy, precision, recall, F1-score

AI – Artificial intelligence; CNN – Convolutional neural network; GNN – Graph neural network; RNN – Recurrent neural network; SVM – Support vector machine; NLP – Natural language processing; ML – Machine learning; AUC – Area under the curve; ICD – International Classification of Diseases; HER – Electronic Health Records; EMR – Electronic medical records; DL – Deep learning

proven effective in enhancing decision-making and patient management, showcasing its adaptability across diverse medical domains.^[32] Similarly, ML algorithms have shown considerable success in diagnostic tasks, such as the prediction and diagnosis of meningitis, as outlined in a recent systematic review.^[33] These examples underscore the versatility of AI and its ability to address complex challenges, reinforcing its potential for broader implementation in automated ICD coding systems. These broader applications emphasize the versatility of AI, which aligns with the future directions identified in this review.

To ensure the practical adoption of AI-based ICD coding, future research should focus on addressing the identified challenges through specific, actionable strategies. Below, the key directions for future development are outlined. Of the 8 included studies, 2 suggested future directions for AI-based automatic ICD coding.^[15,24] After examining the challenges of AI in automatic ICD coding, eight factors require attention and show the need for different future research directions, where significant efforts are needed to develop an automatic ICD coding system. (1) interdisciplinary cooperation;^[14,15,24] (2) data source;^[15] (3) transparency;^[14] (4) exclusive coding of customized research;^[24] (5) national acceptance and coding rules;^[15] (6) reducing the complex problem;^[15,24,25] (7) transfer learning approach for automatic ICD coding;^[15] (8) utilizing active learning and reinforcement learning techniques for clinical classification challenges.^[15]

Future research should explore integrating transfer learning techniques to optimize model performance, especially when working with limited datasets. By leveraging

knowledge from related tasks, transfer learning can significantly reduce the need for large labeled datasets, making models more adaptable to real-world applications. Moreover, implementing explainable AI frameworks can enhance the interpretability of automated ICD coding systems. These frameworks allow the models to provide transparent justifications for their decisions, fostering trust among healthcare providers and enabling more effective collaboration with human coders.

Interdisciplinary cooperation

automated ICD coding models should be developed and improved with input from clinical coders. Algorithms for automated ICD coding should contain important kinds of feedback, such as highlights, corrections, and new rules discovered by human coders. An interface for coders to provide this input ought to be included in automated ICD coding software.^[15,24]

Data source

during the course of our review, we noticed that the absence of publicly accessible benchmark gold standard datasets is one of the primary obstacles to creating an automated ICD coding system.^[15]

Transparency

given that automated ICD coding decisions have an impact on billing and possibly clinical care decisions, transparency is essential. Automated ICD coding models should be easily auditable in order to promote transparent billing practices and contract negotiations. This includes the models' prediction validity, data quality, and reasoning.^[14]

Exclusive coding of customized research

depending on the kind of codes that an automated ICD coding system is intended to produce – such as billing and research – it might satisfy a variety of requirements. In order to forecast diagnosis-related groups, which affect service billing, wider codes are needed for billing. To track procedures and results, new payment mechanisms such as crowdsourcing and worldwide funding might need more intricate coding. A high degree of granularity is also necessary for research; the whole range of codes can be used more effectively for case detection, phenotyping, and other research-related tasks.

National acceptance and coding rules

a lot of nations, including the US, Canada, and Australia, have their own classification schemes. Certain codes are exclusive to a certain nation. For instance, Australia has a greater variety of spider species than both Canada and the US. Therefore, if the data are coded using distinct classification versions, research investigations carried out in different nations may face an additional barrier. Aside from that, while taking into account the application of coding standards, numerous studies have concentrated on automatic ICD code prediction.^[15]

Reducing the complex problem

we also discovered in our review that some research predicted only 50 or 100 frequent codes, condensed codes to only n digits (3 or 4), or excluded seldom occurring codes from the data to make the problem less complicated. The smaller quantity of reports or uncommon codes in the reports is one of the causes of this. Thus, in order to simplify their work, researchers want to think about utilizing multivariate data or reports.^[15,24,25]

Transfer learning approach for automatic International Classification of Diseases coding

training and testing data are typically taken from the same feature space distribution in ML techniques. However, gathering enough training data to train a model is challenging in the real-world applications. In these situations, transfer learning enables the application of knowledge from a related task to a target one. In biomedical research, this strategy is frequently employed due to its demonstrated effectiveness. A few studies^[34,35] have enhanced classification performance by using transfer learning to ICD-9 automatic coding. Thus, in order to tackle the automatic ICD coding job, researchers could investigate several transfer learning strategies.

Methods of active learning and reinforcement learning for clinical categorization issues

it can be difficult to train a model with few reports, an unbalanced class, or a rare class when utilizing ML or DL techniques. Unbalanced classes,^[36] uncommon classes,^[37] and other biomedical categorization issues have been

found with the use of active learning. Similarly, because reinforcement learning can give bigger incentives to minority classes, it has also been shown to be effective in solving the challenge of classifying unbalanced data.^[38] Given the challenges, time constraints, and costs associated with annotating clinical reports, active learning and reinforcement learning can be particularly useful when large amounts of unlabeled data are readily available.^[39] Therefore, for classification tasks, researchers could explore different approaches within active learning and reinforcement learning to improve efficiency and performance. Additionally, the combination of active learning and reinforcement learning offers a promising avenue to address the issues of imbalanced datasets and rare medical codes. For example, reinforcement learning can reward models for correctly classifying minority classes, enhancing their ability to handle less frequent but clinically significant codes.

Finally, establishing international standards for medical data is critical for ensuring interoperability and consistency across healthcare systems. Unified coding standards and data formats would address the variability in coding practices and facilitate the global adoption of automated ICD coding systems.

In summary, this research demonstrates the promise of automatic coding and classification systems, but it also highlights the necessity for field consideration while using automatic coding. What performance level these systems must have in order to carry out practical real-world clinical activities, such as delivering data for an automated decision support system, a clinical research study, or a quality measurement analysis, is another matter that needs more research. Before we can confidently claim that automated coding and classification systems meet the necessary performance standards for complex clinical coding tasks and are capable of applying relevant guidelines to generate reports, these systems need further refinement. A deeper understanding of the specific tasks for which they will be used is also essential. Future research would benefit from testing these methods on larger datasets, such as those based on ICD-10 or ICD-11, to better assess the performance these models could achieve in contemporary clinical settings. Additionally, addressing the issue of data scarcity by finding ways to integrate available training data from multiple datasets (such as MIMIC-III and CodiEsp) and various ontologies (including ICD-9, ICD-10, and MeSH) could significantly enhance the classification performance of these models. Lastly, it would be fascinating to look into additional applications of the data included in the ICD categorization, such as the usage of hierarchical descriptors as a supplement to the loss function.

Conclusions

Although there are numerous challenges, AI-driven automatic ICD coding holds significant potential. This

review outlines the key challenges, advantages, and future pathways for the broader implementation of AI-based automated ICD coding. The goal of this review was to offer a comprehensive reference and suggest possible directions for future research and the widespread adoption of this technology. Practical implementation of AI-based ICD coding requires several key steps: (1) developing robust data infrastructure to support large-scale data integration and model training, (2) providing comprehensive training programs for healthcare professionals to enhance their ability to work alongside AI systems, and (3) ensuring compliance with local and international regulatory frameworks to guarantee ethical and lawful adoption of AI technologies. By addressing these practical considerations, healthcare systems can effectively integrate AI-driven solutions into real-world clinical environments, enhancing the accuracy and efficiency of medical coding processes.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Kaur R, Ginige JA. Comparative analysis of algorithmic approaches for auto-coding with ICD-10-AM and ACHI. *Stud Health Technol Inform* 2018;252:73-9.
2. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care* 2005;43:480-5.
3. Zhou T, Cao P, Chen Y, Liu K, Zhao J, Niu K, *et al.*, editors. Automatic ICD Coding via Interactive Shared Representation Networks with Self-Distillation Mechanism. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2021.
4. Sarbaz M, Mousavi Baigi SF, Manouchehri Monazah F, Dayani N, Kimiafar K. The trend of normal vaginal delivery and cesarean sections before and after implementing the health system transformation plan based on ICD-10 in the Northeast of Iran: A cross-sectional study. *Health Sci Rep* 2023;6:e1131.
5. Kimiafar K, Baigi SF, Sobhani-Rad D, Ranjbar E, Sarbaz M, Esmaeili M. Knowledge and use of international classification of function, disability and health among rehabilitation specialists of Mashhad university of medical sciences: A cross-sectional study. *Front Health Inform* 2024;13:175.
6. Wang C, Yao C, Chen P, Shi J, Gu Z, Zhou Z. Artificial intelligence algorithm with ICD coding technology guided by embedded electronic medical record system in medical record information management. *Microprocess Microsyst* 2023;98:104962.
7. Lu H, Setiono R, Liu H. Effective data mining using neural networks. *IEEE Trans Knowledge Data Eng* 1996;8:957-61.
8. Heiden-Rootes KM, Salas J, Gebauer S, Witthaus M, Scherrer J, McDaniel K, *et al.* Sexual dysfunction in primary care: An exploratory descriptive analysis of medical record diagnoses. *J Sex Med* 2017;14:1318-26.
9. Neville TH, Tarn DM, Yamamoto M, Garber BJ, Wenger NS. Understanding factors contributing to inappropriate critical care: A mixed-methods analysis of medical record documentation. *J Palliat Med* 2017;20:1260-6.
10. Siregar R. Performance analysis of AES-Blowfish hybrid algorithm for security of patient medical record data. *J Phys Conf Ser*. 2018;1007:012018.
11. National Center for Health Statistics. International Classification of Diseases, (ICD-10-CM/PCS) Transition – Background; 2015. Available from: https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm. [Last accessed on 2024 Jan 01].
12. Campbell S, Giadresco K. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *Health Inf Manage J* 2020;49:5-18.
13. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol*. 2017;2:e000101.
14. Moons E, Khanna A, Akkasi A, Moens MF. A comparison of deep learning methods for ICD coding of clinical records. *Appl Sci* 2020;10:5262.
15. Kaur R, Ginige JA, Obst O. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Syst Appl* 2023;213:118997.
16. Ghaddaripouri K, Baigi SF, Noori N, Habibi MR. Investigating the effect of virtual reality on reducing the anxiety in children: A systematic review. *Front Health Inform* 2022;11:114.
17. Mousavi Baigi SF, Sarbaz M, Sobhani-Rad D, Kimiafar K. A comparative study of rehabilitation information systems in 8 countries: A literature review. *Iran Rehabil J* 2023;21:1-16.
18. Kimiafar K, Sarbaz M, Tabatabaei SM, Ghaddaripouri K, Mousavi AS, Mehneh MR, *et al.* Artificial intelligence literacy among healthcare professionals and students: A systematic review. *Front Health Inform* 2023;12:168.
19. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: A systematic review. *Health Sci Rep* 2023;6:e1138.
20. Baigi S, Mehneh M, Sarbaz M, Aval R, Kimiafar K. Telerehabilitation in response to critical coronavirus: A systematic review based on current evidence. *J Isfahan Med Sch*. 2022;40:498-508.
21. Yan C, Fu X, Liu X, Zhang Y, Gao Y, Wu J, *et al.* A survey of automated International Classification of Diseases coding: Development, challenges, and applications. *Intell Med* 2022;2:161-73.
22. Ji S, Li X, Sun W, Dong H, Taalas A, Zhang Y, *et al.* A unified review of deep learning for automated medical coding. *ACM Comput Surv*. 2024;56:Article 306.
23. Li X, Zhao X, Zhang Y, Xing C, editors. Towards Automatic ICD Coding via Knowledge Enhanced Multi-Task Learning. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management; 2023.
24. Venkatesh KP, Raza MM, Kvedar JC. Automating the overburdened clinical coding system: Challenges and next steps. *NPJ Digit Med* 2023;6:16.
25. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc* 2010;17:646-51.
26. Wallace K, Masud JH. Artificial intelligence for prediction of International Classification of Disease codes. *Bangabandhu Sheikh Mujib Med Univ J*. 2023;16:118-23.
27. Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, *et al.* Automated clinical coding: What, why, and where we are? *NPJ Digit Med* 2022;5:159.
28. Critical Appraisal Skills Programme (CASP). CASP Checklists.

- Oxford: CASP; 2018. Available from: <https://casp-uk.net/casp-tools-checklists>. [Last accessed on 2024 Jan 01].
29. Pascual D, Luck S, Wattenhofer R. Proceedings of the 20th Workshop on Biomedical Language Processing. Online: Association for Computational Linguistics; 2021. p. 54-63.
 30. Dai HJ, Wang CK, Chen CC, Liou CS, Lu AT, Lai CH, *et al.* Evaluating a natural language processing-driven, AI-assisted international classification of diseases, 10th revision, clinical modification, coding system for diagnosis related groups in a real hospital environment: Algorithm development and validation study. *J Med Internet Res* 2024;26:e58278.
 31. Mousavi Baigi SF, Dahmardeh Kemmak F, Sarbaz M, Norouzi Aval R, Kimiafar K. Application of artificial intelligence in occupational therapy. *Health Educ Health Promot* 2024;12:513-20.
 32. Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Mousavi Baigi SF, Rezaei Sarsari M, Dahmardeh Kemmak F, *et al.* The effect of machine learning algorithms in the prediction, and diagnosis of meningitis: A systematic review. *Health Sci Rep* 2024;7:e1893.
 33. Luo J, Xiao C, Glass L, Sun J, Ma F, editors. Fusion: towards automated ICD coding via feature compression. In: Findings of the Association for Computational Linguistics. Online: ACL-IJCNLP; 2021. p. 2096-101.
 34. Zeng M, Li M, Fei Z, Yu Y, Pan Y, Wang J. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing* 2019;324:43-50.
 35. Rios A, Kavuluru R. Neural transfer learning for assigning diagnosis codes to EMRs. *Artif Intell Med* 2019;96:116-22.
 36. Hospedales TM, Gong S, Xiang T. Finding rare classes: Active learning with generative and discriminative models. *IEEE Trans Knowl Data Eng* 2011;25:374-86.
 37. Flores CA, Figueroa RL, Pezoa JE. Active learning for biomedical text classification based on automatically generated regular expressions. *IEEE Access* 2021;38767:77-9.
 38. Lin E, Chen Q, Qi X. Deep reinforcement learning for imbalanced classification. *Appl Intell* 2020;50:2488-502.
 39. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, *et al.* Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst Appl* 2019;116:494-520.